Department of Defense

High Performance Computing Modernization Program

Cluster Computing Experiences, Performance Measurements & Valuation

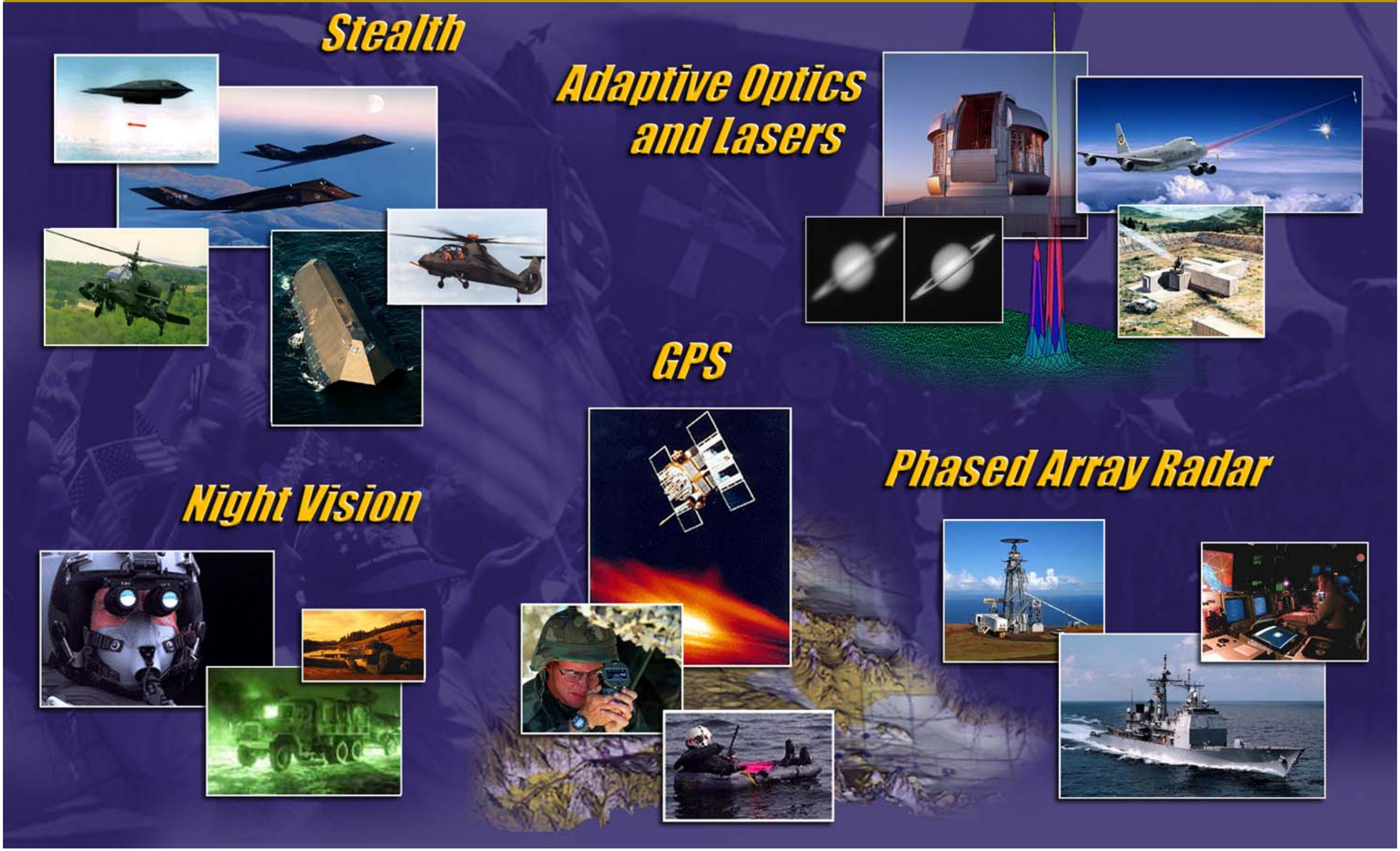
Mr. Cray Henry, Director

http://www.hpcmo.hpc.mil

Agenda

- High Performance Computing Modernization Program Overview
- Valuation & Performance Measurement
- HPCMP and Commodity Cluster Computing Experiences
- End Notes

A Focus on Revolutionary Advances

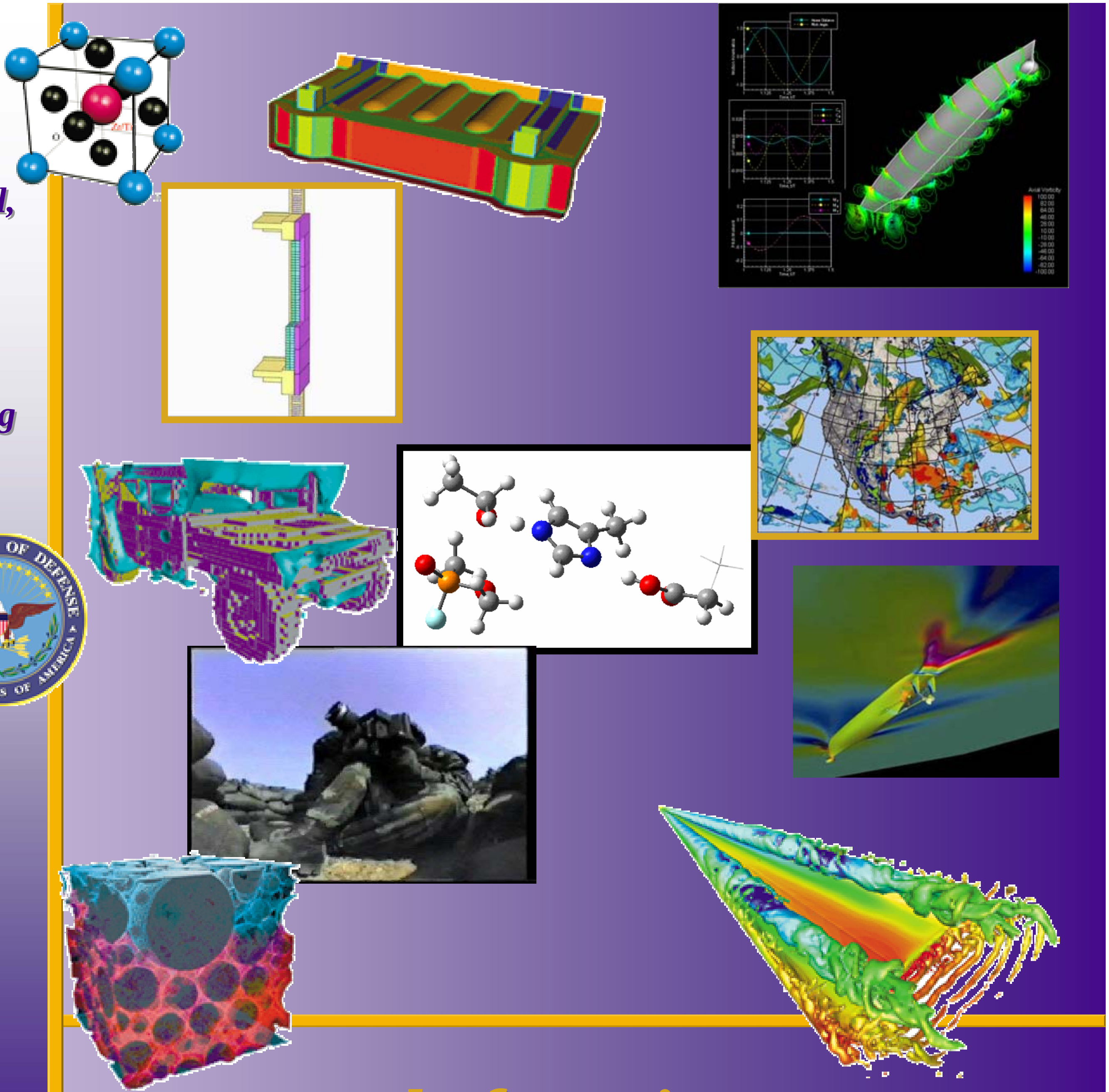


Mission

Deliver world-class commercial, high-end, high performance computational capability to the DoD's science and technology (S&T) and test and evaluation (T&E) communities, facilitating the rapid application of advanced technology into superior warfighting capabilities.

Vision

A pervasive culture existing among DoD’s scientists and engineers where they routinely use advanced computational environments to solve the most demanding problems.

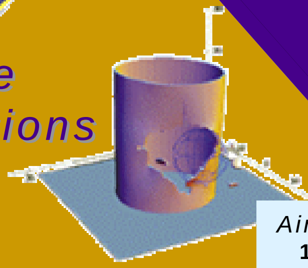


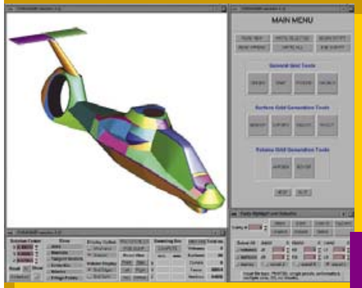
Tools for Discovery

HPC Modernization Program

Software Applications Support

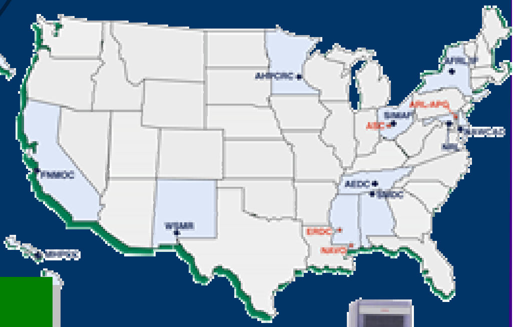
CHSSI

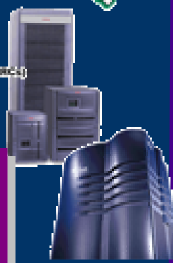




Warfighter Support

HPC Centers


AHPCRC
ASC
SIMAF
AFRL/IF
ARL-APG
FNMOC
WSMR
AEDC
SMDC
ERDC
NAVO
Major Shared Resource Centers (MSRCs)
Distributed Centers (DCs)

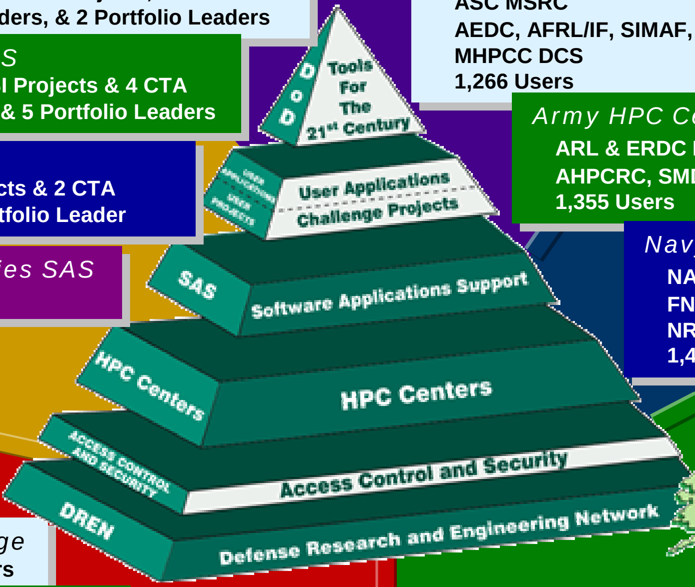


Air Force SAS
14 CHSSI Projects, 3 CTA Leaders, & 2 Portfolio Leaders

Army SAS
11 CHSSI Projects & 4 CTA Leaders & 5 Portfolio Leaders

Navy SAS
17 CHSSI Projects & 2 CTA Leaders& 1 Portfolio Leader

Defense Agencies SAS
1 CHSSI


DoD
Tools For The 21st Century
User Applications
DREN Projects
User Applications
Challenge Projects
SAS
Software Applications Support
HPC Centers
HPC Centers
ACCESS CONTROL AND SECURITY
Access Control and Security
DREN
Defense Research and Engineering Network

Air Force HPC Centers
ASC MSRC
AEDC, AFRL/IF, SIMAF, & MHPCC DCS
1,266 Users

Army HPC Centers
ARL & ERDC MSRC
AHPCRC, SMDC, & WSMR DCs
1,355 Users

Navy HPC Centers
NAVO MSRC
FNMOC, NAWCAD & NRL DCs
1,437 Users

Defense Agencies
DARPA, DTRA, JNIC, MDA, & OTE
285 Users

PET

Resource Mgmt

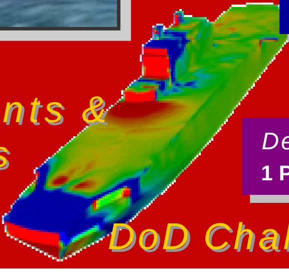
Air Force Challenge
10 Proj. and 1,286 Users

Army Challenge
11 Proj. and 1,335 Users

Navy Challenge
16 Proj. and 1,437 Users

Army & Navy Challenge
1 Proj.

Defense Agencies Challenge
1 Proj.



Requirements & Allocations

DoD Challenges

DREN

Air Force DREN
12 Sites

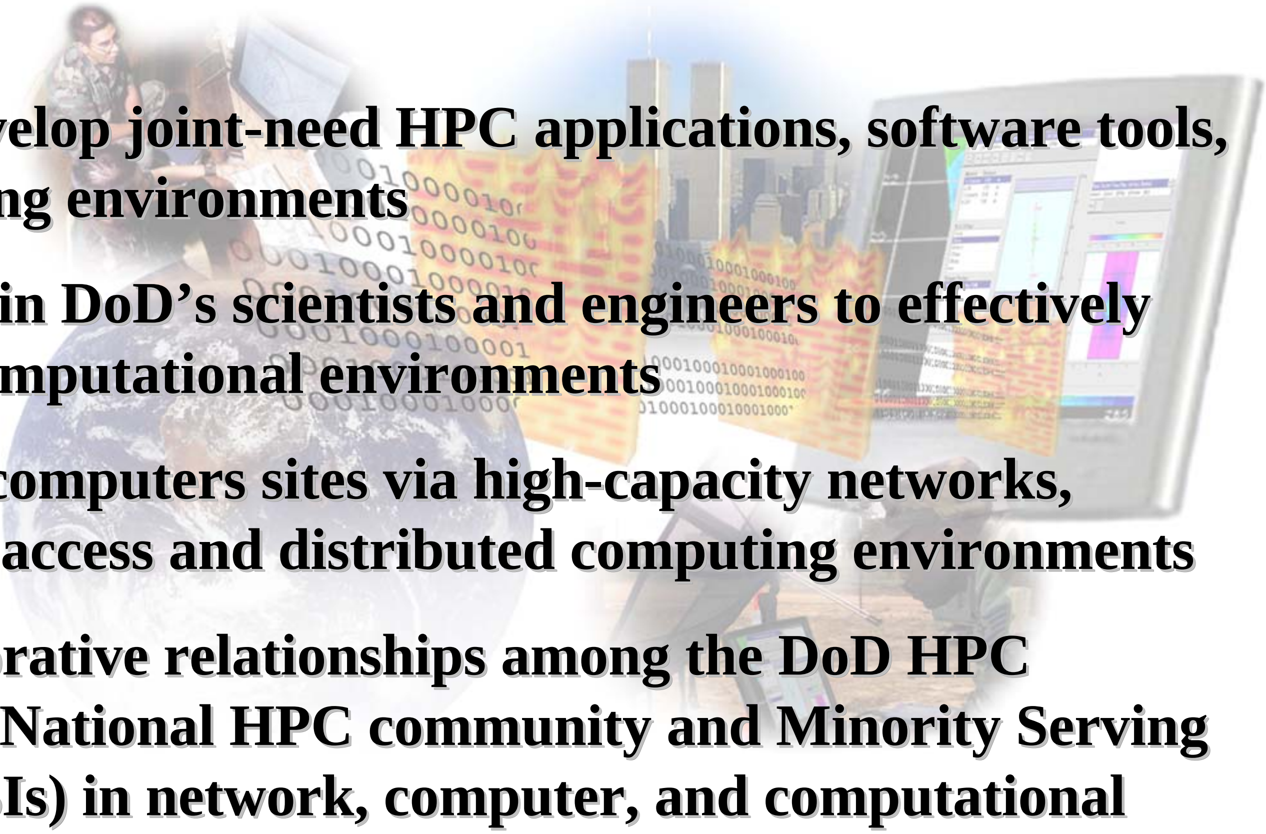
Army DREN
25 Sites

Navy DREN
25 Sites

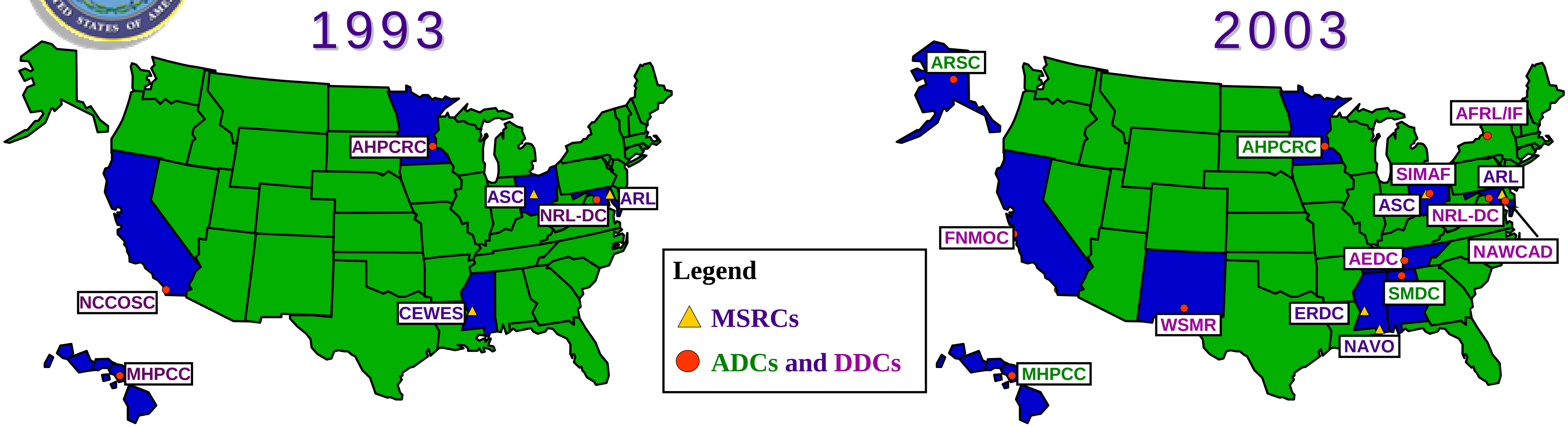
Defense Agencies DREN
7 Sites

HPCMP Goals

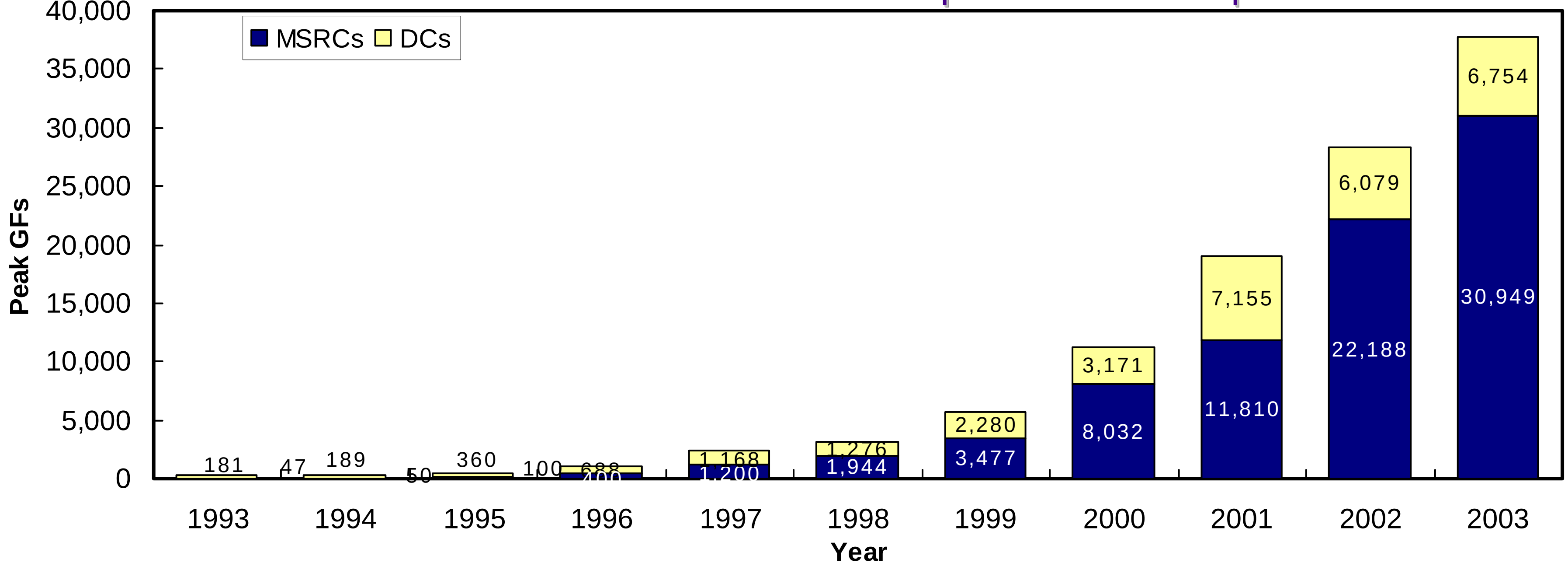
- Provide the best commercially available high-end HPC capability
- Acquire and develop joint-need HPC applications, software tools, and programming environments
- Educate and train DoD's scientists and engineers to effectively use advanced computational environments
- Link users and computers sites via high-capacity networks, facilitating user access and distributed computing environments
- Promote collaborative relationships among the DoD HPC community, the National HPC community and Minority Serving Institutions (MSIs) in network, computer, and computational science



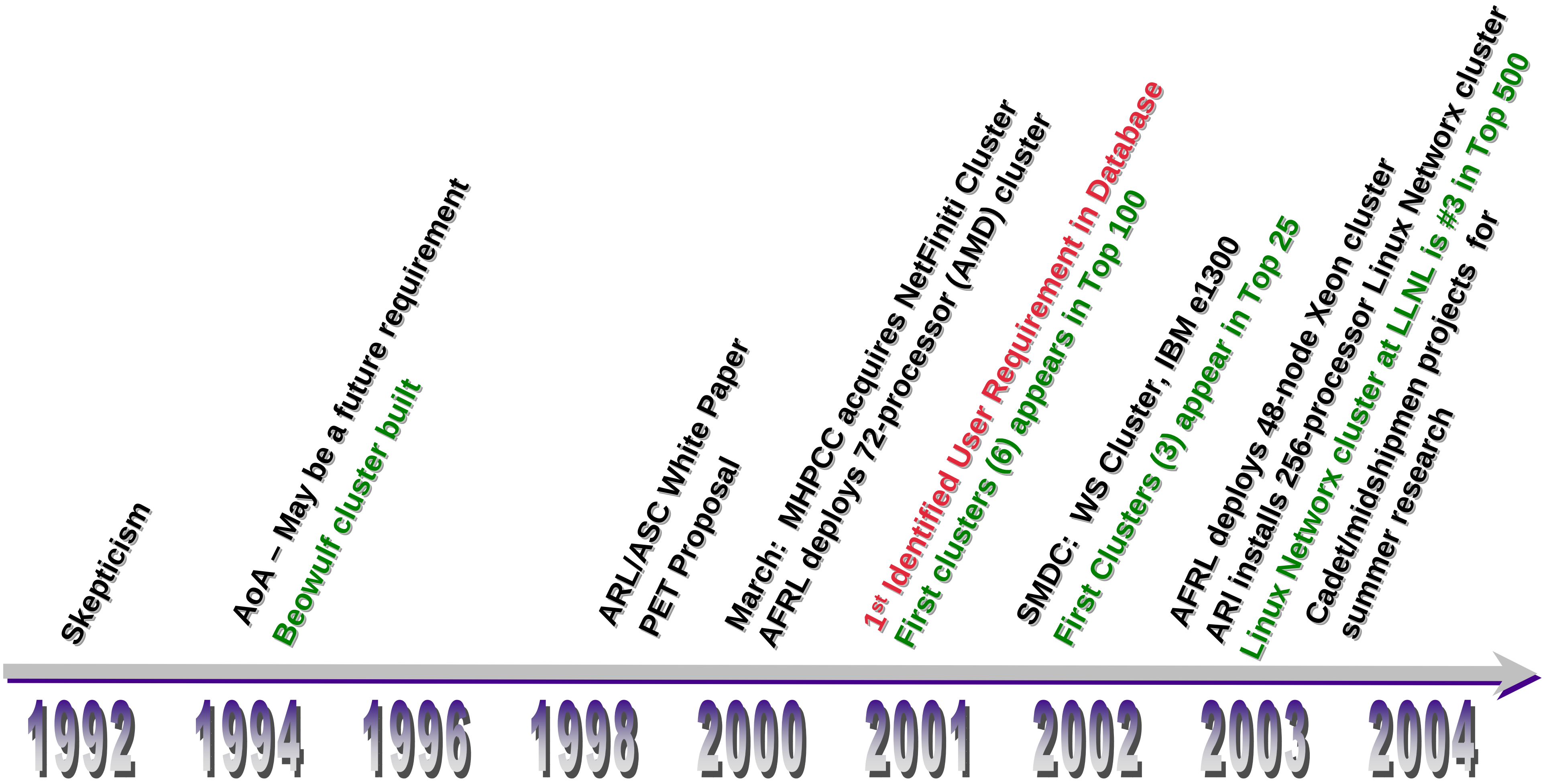
HPCMP Centers



Total HPCMP End-of-Year Computational Capabilities



HPCMP and Commodity Cluster Computing



Skepticism

AoA – May be a future requirement

Beowulf cluster built

ARL/ASC White Paper

PET Proposal

March: MHPCC acquires NetFiniti Cluster

AFRL deploys 72-processor (AMD) cluster

1st Identified User Requirement in Database

First clusters (6) appears in Top 100

SMDC: WS Cluster, IBM e1300

First Clusters (3) appear in Top 25

AFRL deploys 48-node Xeon cluster

ARI installs 256-processor Linux Networx cluster

Linux Networx cluster at LLNL is #3 in Top 500

Cadet/midshipmen projects for summer research

Intense Interest on Clusters

- Top 500 List identifies 149 clusters

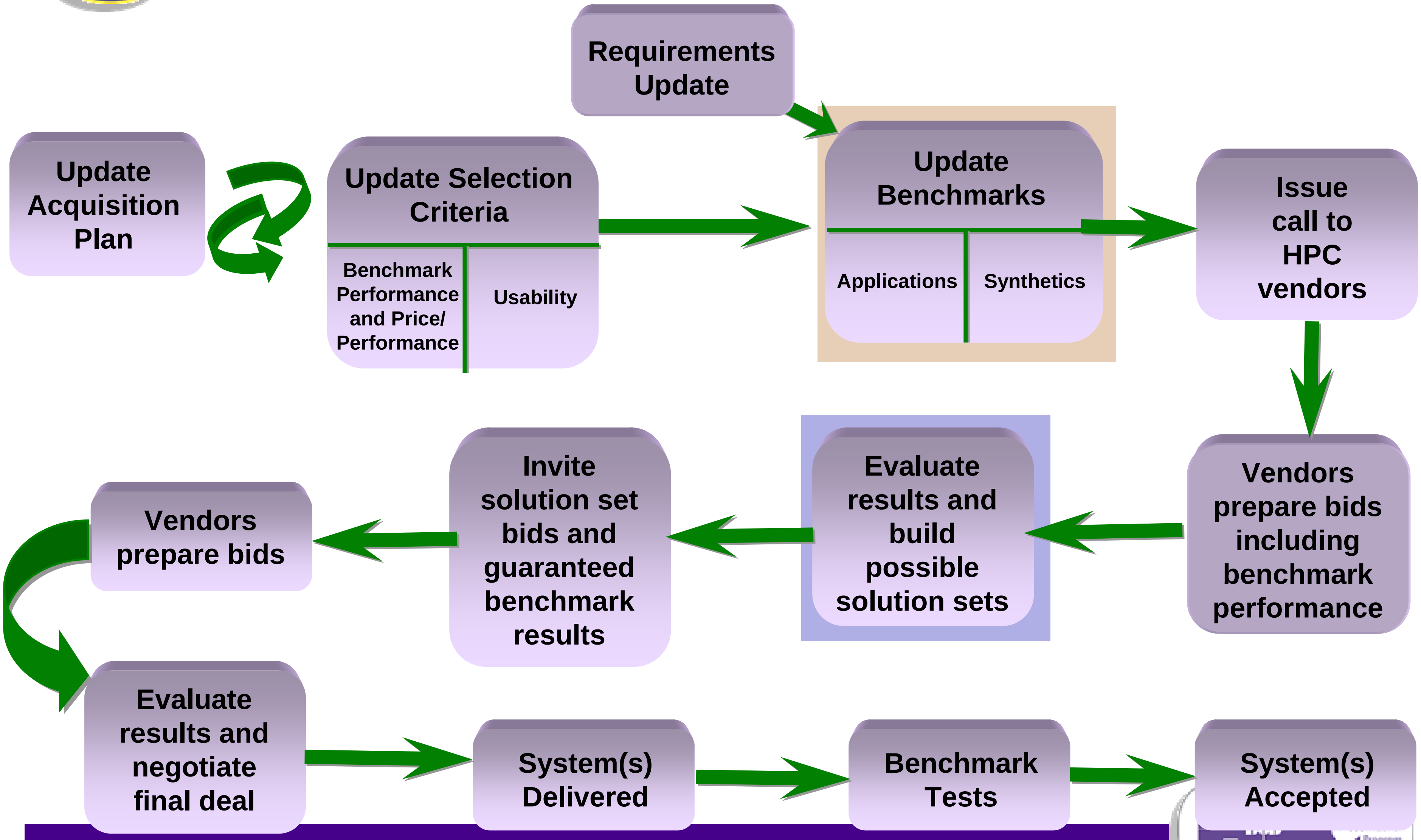
- Grid Computing

But what is the Real Performance of clusters on real workloads?

Technology Insertion-XX

- Purpose of TI-XX
  - Buy Systems Based Upon User Requirements
  - Focus on Program-wide Acquisition Strategy
  - Determine Program-wide Best Value
- How
  - Evaluate Performance, Price/Performance and Usability of Multiple OEMS, Using Benchmarks and Qualitative Assessments Based on User and Operator Needs

Technology Insertion (TI) Flow Chart



Requirements Update

Update Acquisition Plan

Update Selection Criteria

Benchmark Performance and Price/ Performance

Usability

Update Benchmarks

Applications

Synthetics

Issue call to HPC vendors

Vendors prepare bids including benchmark performance

Evaluate results and build possible solution sets

Invite solution set bids and guaranteed benchmark results

Vendors prepare bids

Evaluate results and negotiate final deal

System(s) Delivered

Benchmark Tests

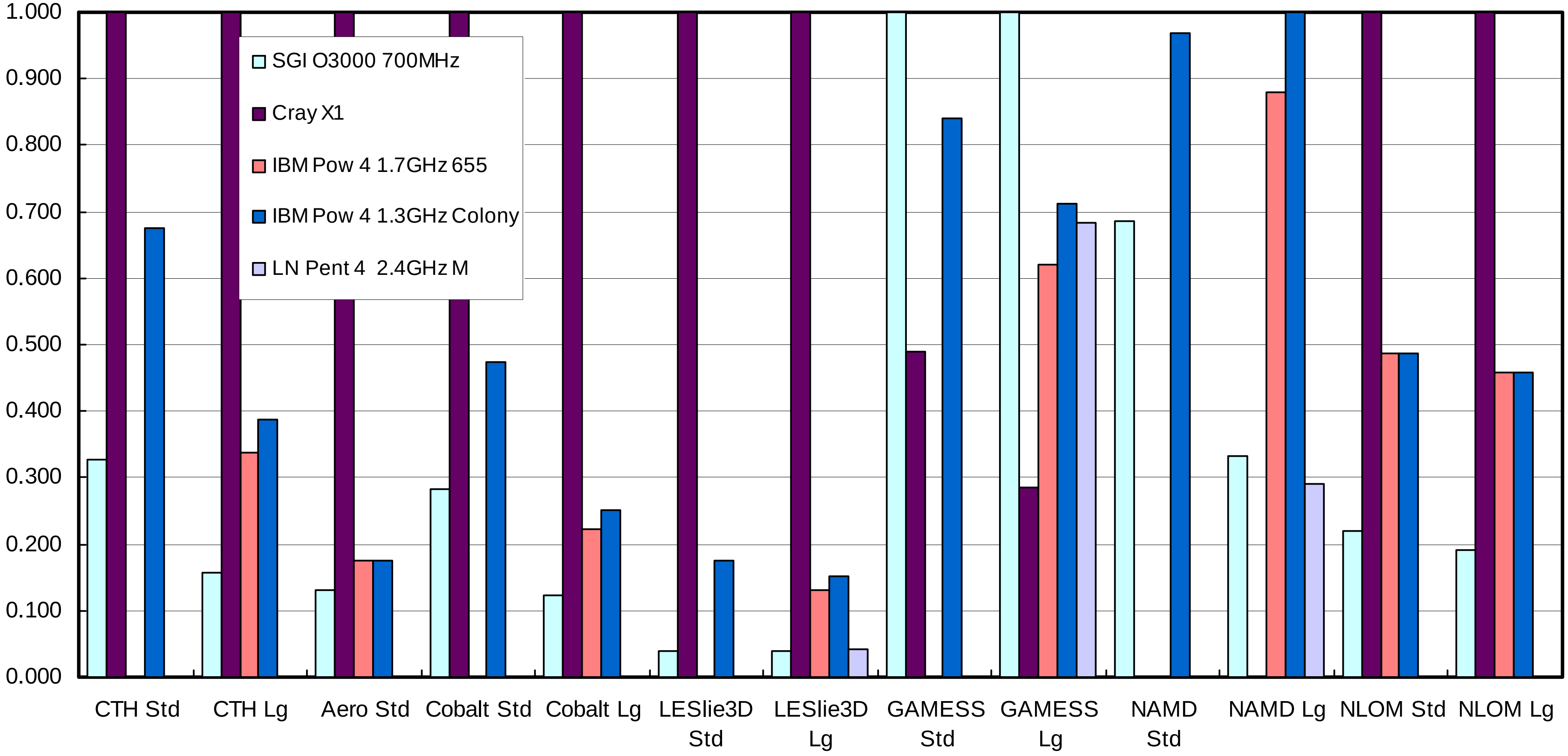
System(s) Accepted

Emphasis on Performance
Time to Solution

- Establish a DoD standard benchmark time for each application benchmark case
  - NAVO IBM SP P3 chosen as standard DoD system
- Benchmark timings (at least three on each test case) are requested for systems that meet or beat the DoD standard benchmark times by at least a factor of two (preferably four)
- Benchmark timings may be extrapolated provided they are guaranteed, but at least one actual timing must be provided

HPC System Performance Results

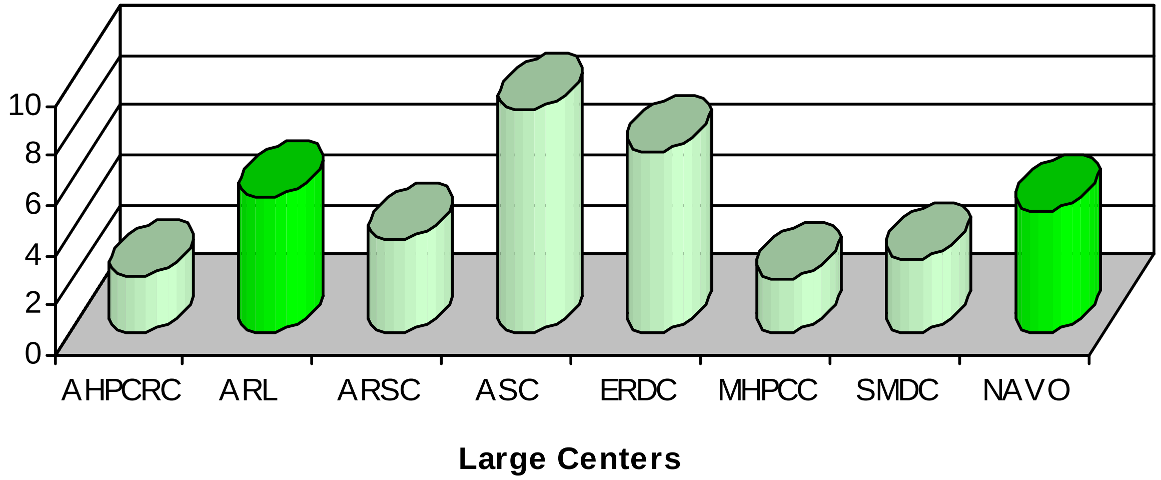
Normalized Capability Performance Scores



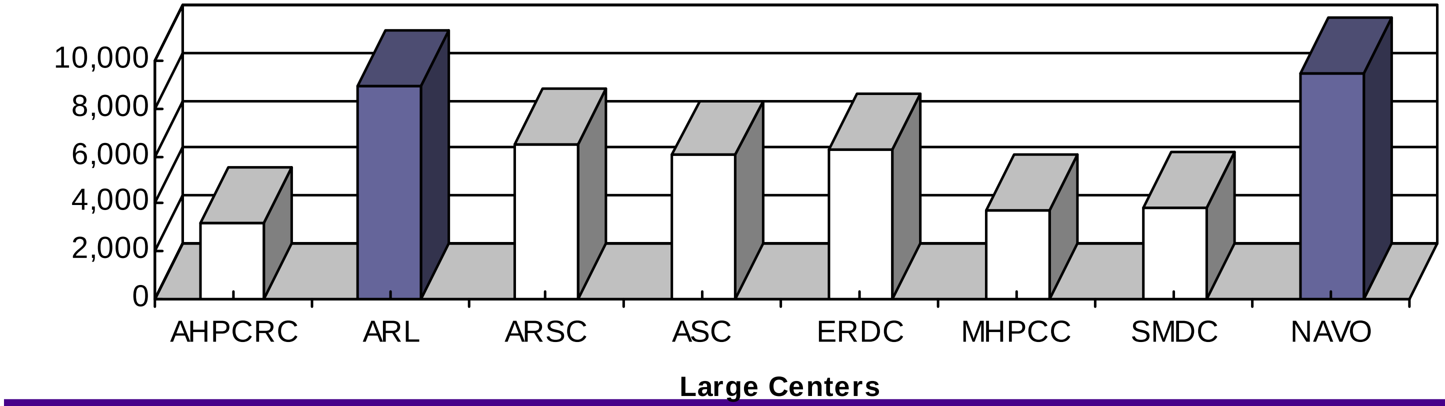
Capturing True Performance

Benchmarks

Capacity of MSRCs in Habu Equivalents



Capacity of MSRCs in Peak GFlop-years



Top 500 or Peak G-Flops is not a Measure of Real Performance

Solution Set Building

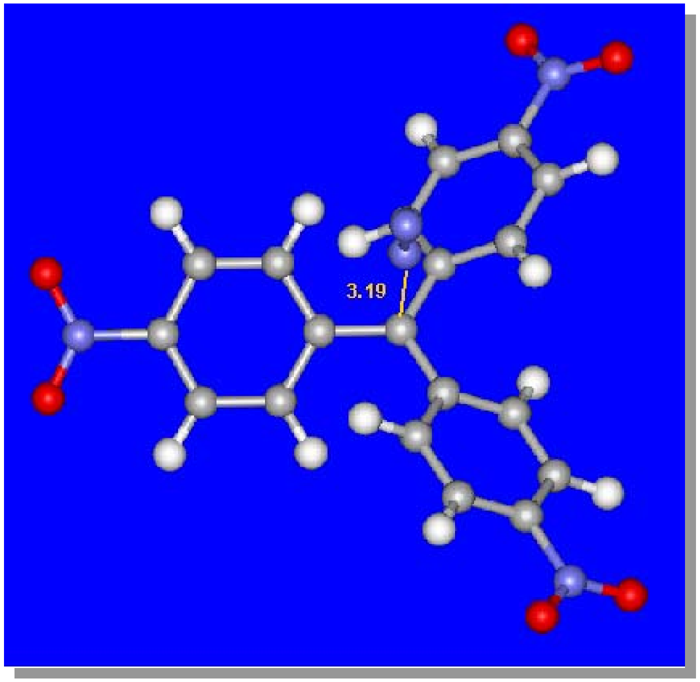
System					CTH Std	CTH Lg	Aero	Cobalt S	Cobalt L
Unclassified Benchmark Weights =					5.53%	3.35%	10.94%	8.20%	12.68%
Classified Benchmark Weights =					XX	XX	XX	XX	XX
System	# Proc	Number	Cost($M)	Total					
Cray X1	128	0	$1	$0.000	0.000	0.000	0.000	0.000	0.000
Cray X1	64	0	$1	$0.000	0.000	0.000	0.000	0.000	0.000
Cray X1	256	0	$1	$0.000	0.000	0.000	0.000	0.000	0.000
IBM Pw 4 1.7GHz 655	512	0	$1	$0.000	0.000	0.000	0.000	0.000	0.000
IBM Pw 4 1.7GHz 690	160	0	$1	$0.000	0.000	0.000	0.000	0.000	0.000
IBM Pw 4 1.7GHz 690	128	0	$1	$0.000	0.000	0.000	0.000	0.000	0.000
LN Pent 4 2.4GHz Q	512	0	$1	$0.000	0.000	0.000	0.000	0.000	0.000
LN Pent 4 2.4GHz Q	256	0	$1	$0.000	0.000	0.000	0.000	0.000	0.000
LN Pent 4 2.4GHz M	512	0	$1	$0.000	0.000	0.000	0.000	0.000	0.000
LN Pent 4 2.4GHz M	256	0	$1	$0.000	0.000	0.000	0.000	0.000	0.000
SGI O3000 600MHz	256	0	$1	$0.000	0.000	0.000	0.000	0.000	0.000
SGI O3000 700MHz	1024	0	$1	$0.000	0.000	0.000	0.000	0.000	0.000
SGI O3000 700MHz	512	0	$1	$0.000	0.000	0.000	0.000	0.000	0.000
SGI O3000 700MHz	256	0	$1	$0.000	0.000	0.000	0.000	0.000	0.000
SGI O3000t 700MHz	256	0	$1	$0.000	0.000	0.000	0.000	0.000	0.000
Total for Alternative				$0.000	0.000	0.000	0.000	0.000	0.000
Application Percentage					0.000%	0.000%	0.000%	0.000%	0.000%

Total Performance Score

□

New Materials Design

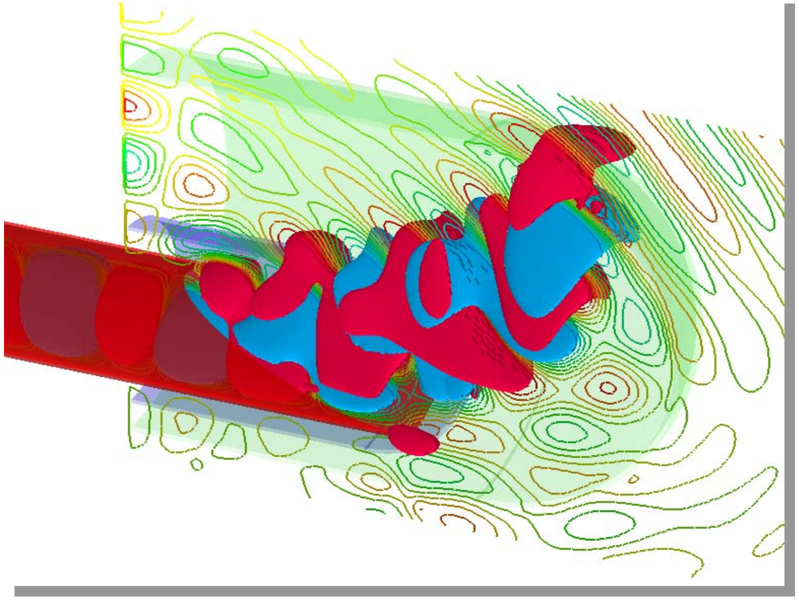
Platform(s)	Location		CPU Resources (processor-hours)	
	First Choice	Second Choice	Request	Minimum Acceptable
Cray T3E	NAVO	ERDC	120,000	90,000
Cray T3E	ERDC	NAVO	300,000	250,000
Linux Cluster	MHPCC	n/a	180,000	160,000
Compaq ES40	ASC	n/a	150,000	125,000
Compaq GS320	ASC	n/a	150,000	125,000
IBM SP	MHPCC	ASC	300,000	260,000
IBM-SP/P3	ASC	n/a	150,000	125,000
IBM SP/P3	ASC	MHPCC	60,000	40,000
IBM SP/P3	ARSC	ARL	40,000	30,000
Cray SV1	ARSC	NAVO	2,000	1,000



Major Applications Software:
GAMESS (CHSSI), FMD (CHSSI), CMD (CHSSI), Gaussian98.

Virtual Prototyping of Directed Energy Weapons

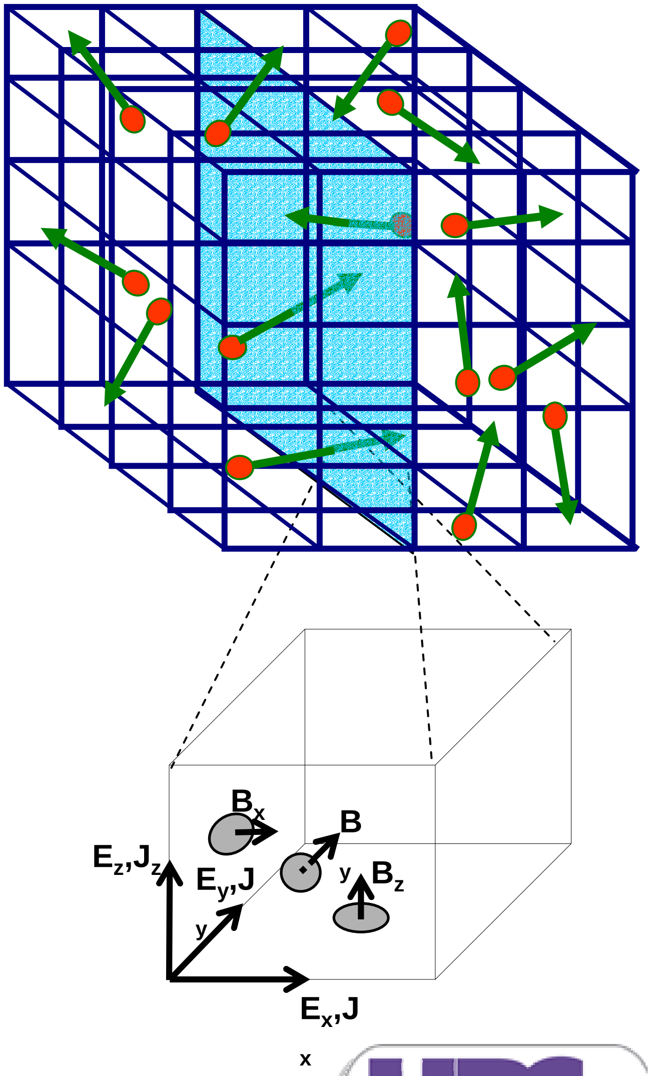
	Location		CPU Resources (processor-hours)	
Platform(s)	First Choice	Second Choice	Request	Minimum Acceptable
IBM SP P3	ARL	NAVO	500,000	400,000
Compaq SC40/45	ERDC	ASC	300,000	250,000
IBM Netfinity	MHPCC		100,000	75,000



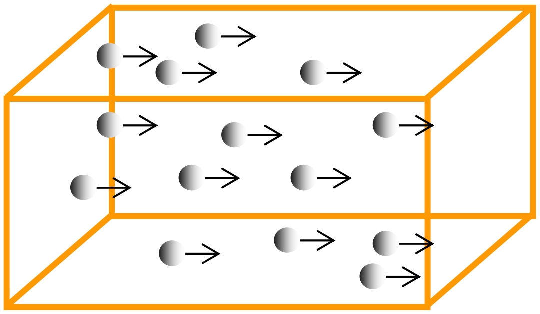
Major Application Software:
ICEPIC

ICEPIC

- *ICEPIC* is a beam-plasma physics electromagnetic particle-in-cell code that solves Maxwell's equations and the relativistic Lortntz force law
- Written in ANSI standard C with MPI to be portable to all Unix or Linux platforms
- Compiled with GCC –03 optimization
- MPICH 2.4



ICEPIC Test Problem Descriptions



- Memory requirements for data structures
  - Cell: 256 Bytes; Particle: 48 Bytes
- Typical application problem has both:
  - cell-dominated regions (>10 cells/particle), and
  - particle-dominated regions (>10 particles/cell)
- Two test problems designed to investigate both limits:
  - 3 dimensional box with square cross-section
    - Cell-Dominated
      - » ≈1 million cells; 1,000 particles (requires ≈ 256 MB memory)
    - Particle-Dominated
      - » 50,000 cells; 8 million particles (requires 390 MB memory)
  - In both cases, data fits into memory on 1 processor for all platforms

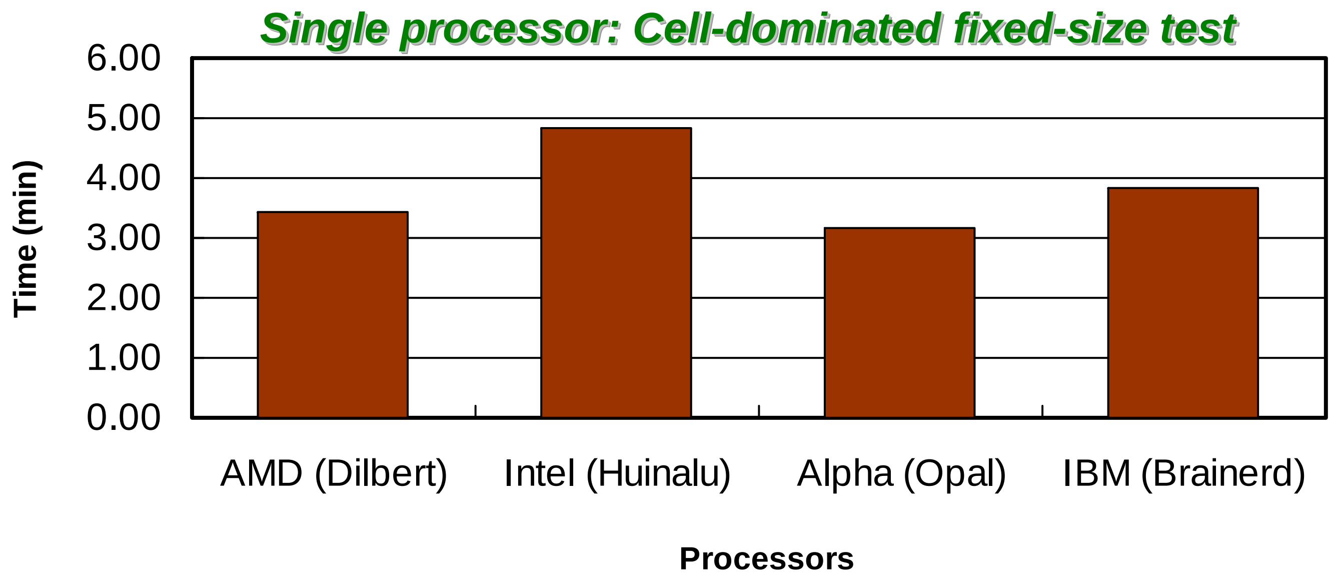
Clusters Used

- AFRL Custom-made LINUX Cluster “Dilbert”
  - 18 nodes; 2 processors/node; 2 GB memory/node
    - 36 AMD Athlon cpu-s
      - 1.6 GHz scalar
  - Red Hat Linux 7.1 with 2.4.19 kernel
  - Nodes connected via 100 Mbit/s Ethernet from a single switch
- MHPCC ADC LINUX Cluster “Huinalu”
  - 256 nodes; 2 processors/node; 1 GB memory/node
    - 512 Intel Pentium III cpu-s
      - 933 MHz scalar
  - Red Hat Linux with 2.4.18 kernel
  - Nodes connected via two options:
    - 100 Mbit/s Ethernet
    - 200 MByte/s Myrinet

Non-Clusters Used

- ARL MSRC “Brainerd”
  - 32 nodes; 16 processors/node; 16 GB memory/node
    - 512 IBM SP-P3 cpu-s
      - 375 MHz superscalar (2 mults and 2 adds per cycle)
  - Nodes connected via 32-port 200 MByte/s Colony switch
- ERDC MSRC “Opal”
  - 128 nodes; 4 processors/node; 4 GB memory/node
    - 512 DEC Alpha EV 68 cpu-s
      - 833 MHz superscalar (1 mult and 1 add per cycle)
  - Nodes connected via 64-port, single-rail 200 MByte/s Quadrics switch

Serial Performance of Component Processors



AMD –
1.6GHz Athlon

Intel –
933 MHz P3

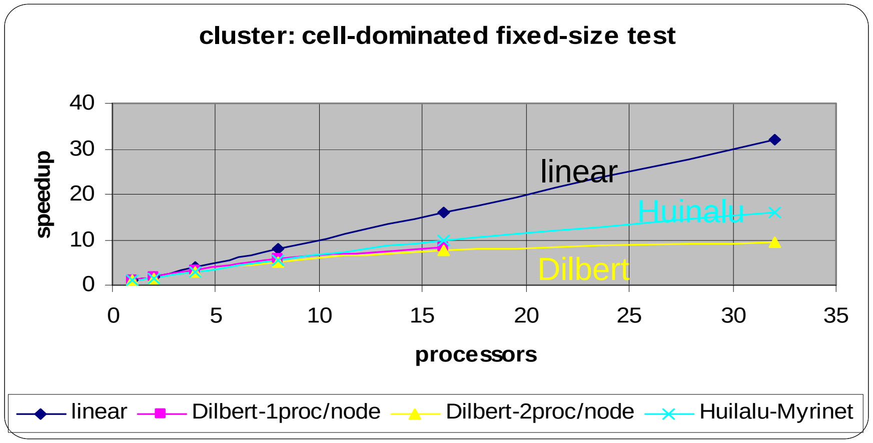
Alpha –
833MHz EV68

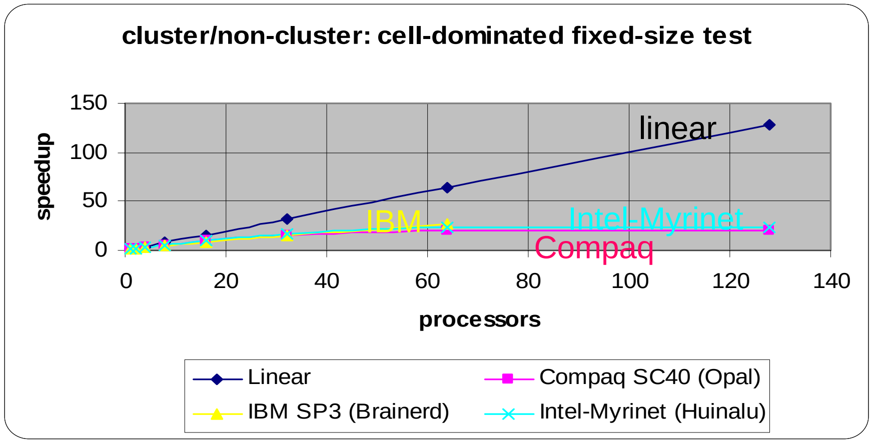
IBM –
375MHz Pw3

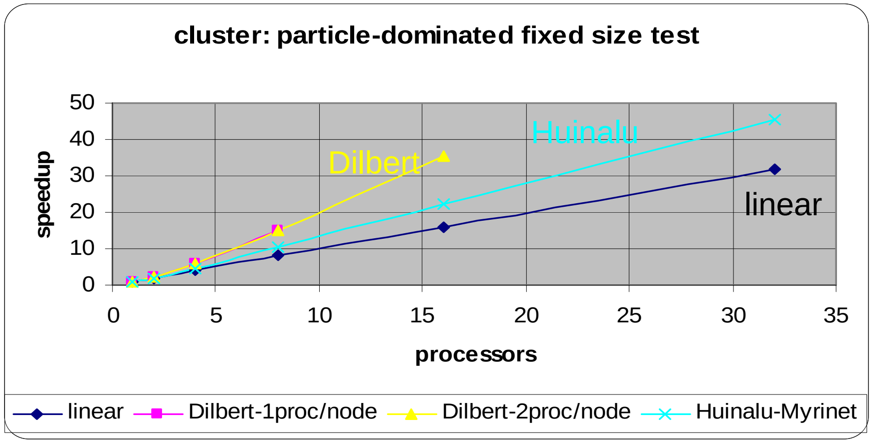
For these *ICEPIC* cell-dominated simulations:

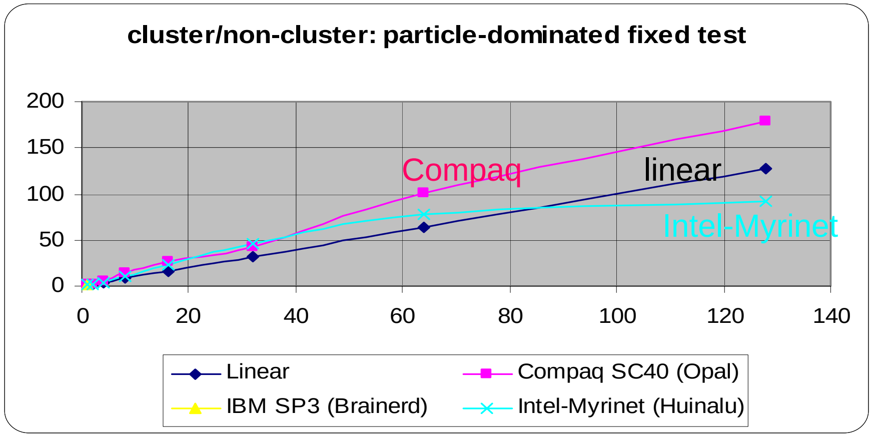
- ***Circa* 2002 Dilbert (1.6 GHz AMD Athlon) processor outperforms the *circa* 2000 Huinalu (933 MHz IBM Pentium III) processor and *circa* 1999 Brainerd (375 MHz IBM SP-P3) processor**
- **Dilbert (1.6 GHz AMD Athlon) processor performs comparably to Opal (833 MHz DEC Alpha) processor**

Parallel Performance:
Speedup of Fixed-Sized Problem


cluster: cell-dominated fixed-size test

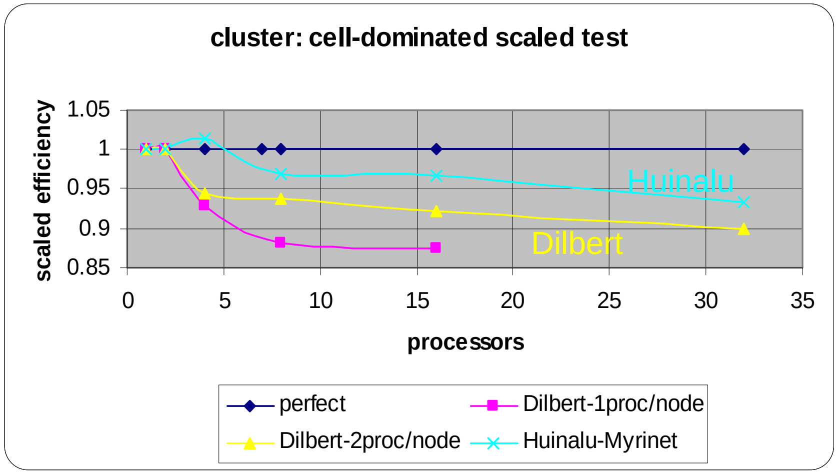

cluster/non-cluster: cell-dominated fixed-size test

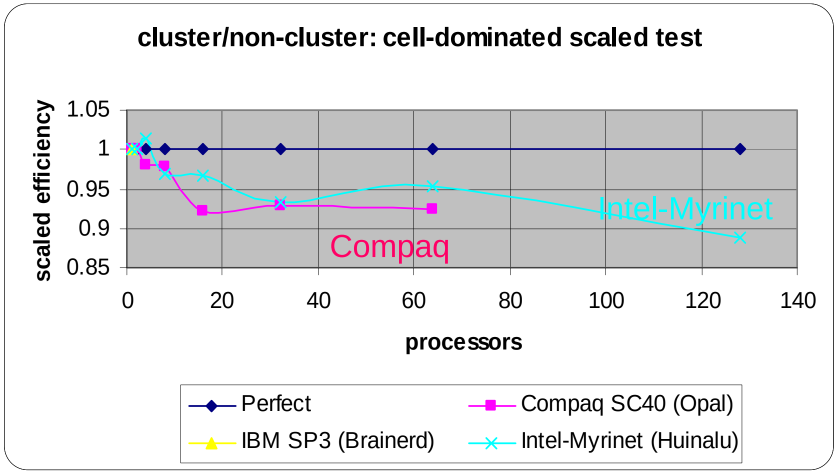

cluster: particle-dominated fixed size test

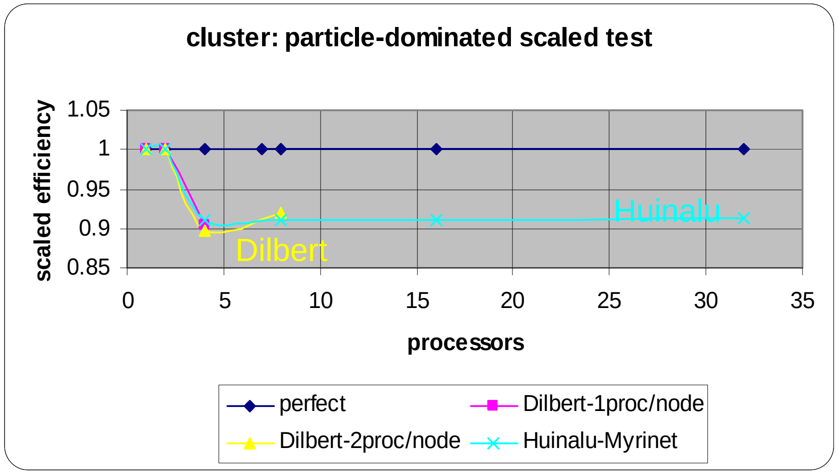

cluster/non-cluster: particle-dominated fixed test

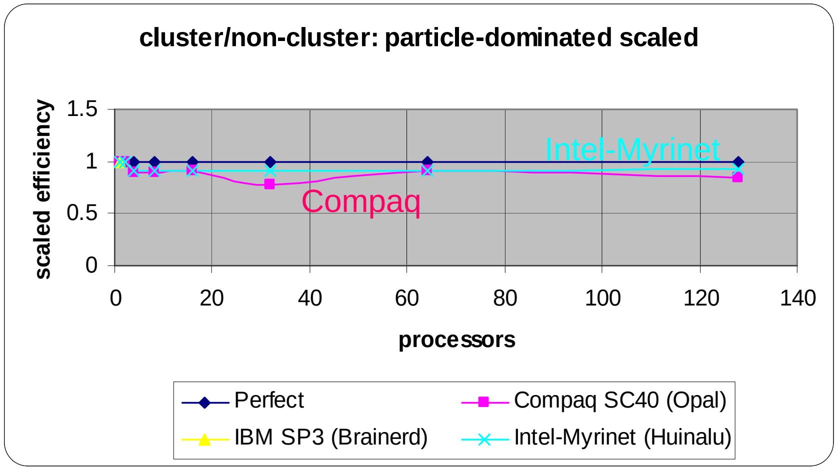
- Super-linear speedup for particle-dominated tests is a consequence of large number of particles looking up small amount of cell data.
- As the number of processors increases, more cell data fits into cache

Parallel Performance:
Efficiency of Scaled Problem


cluster: cell-dominated scaled test


cluster/non-cluster: cell-dominated scaled test


cluster: particle-dominated scaled test


cluster/non-cluster: particle-dominated scaled

Cluster Performance Observations

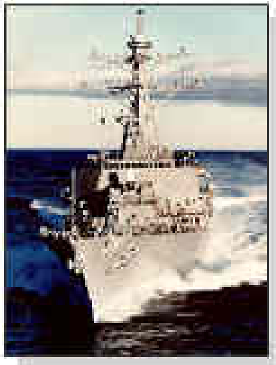
- Reliability and Reproducibility of parallel run results
  - Data presented is “best case,” not average
  - 100 Mbit/s Intel/AMD Ethernet (Huinalu and Dilberts)
    - For all numbers of processors:
      - » Runs always get through the queue
      - » Timings are reproducible to within 3%
  - 200 MByte/s Intel Myrinet (Huinalu)
    - For up to 64 processors:
      - » Runs always get through the queue
      - » Timings are reproducible to within 5%
    - For more than 64 processors:
      - » Runs get through the queue about half the time
      - » Timings vary by up to 40%

Non-cluster Performance Observations

- Reliability and Reproducibility of parallel run results
  - Data presented is "best case," not average
  - 200 MByte/s Compaq Quadrics (Opal)
    - For all numbers of processors:
      - Runs usually get through the que in a timely fashion
      - Timings are reproducible to within 5%
  - 200 MByte/s IBM Colony (Brainerd)
    - For all numbers of processors:
      - It usually takes a long time for runs to get through the que
      - Timings are reproducible to within 5%

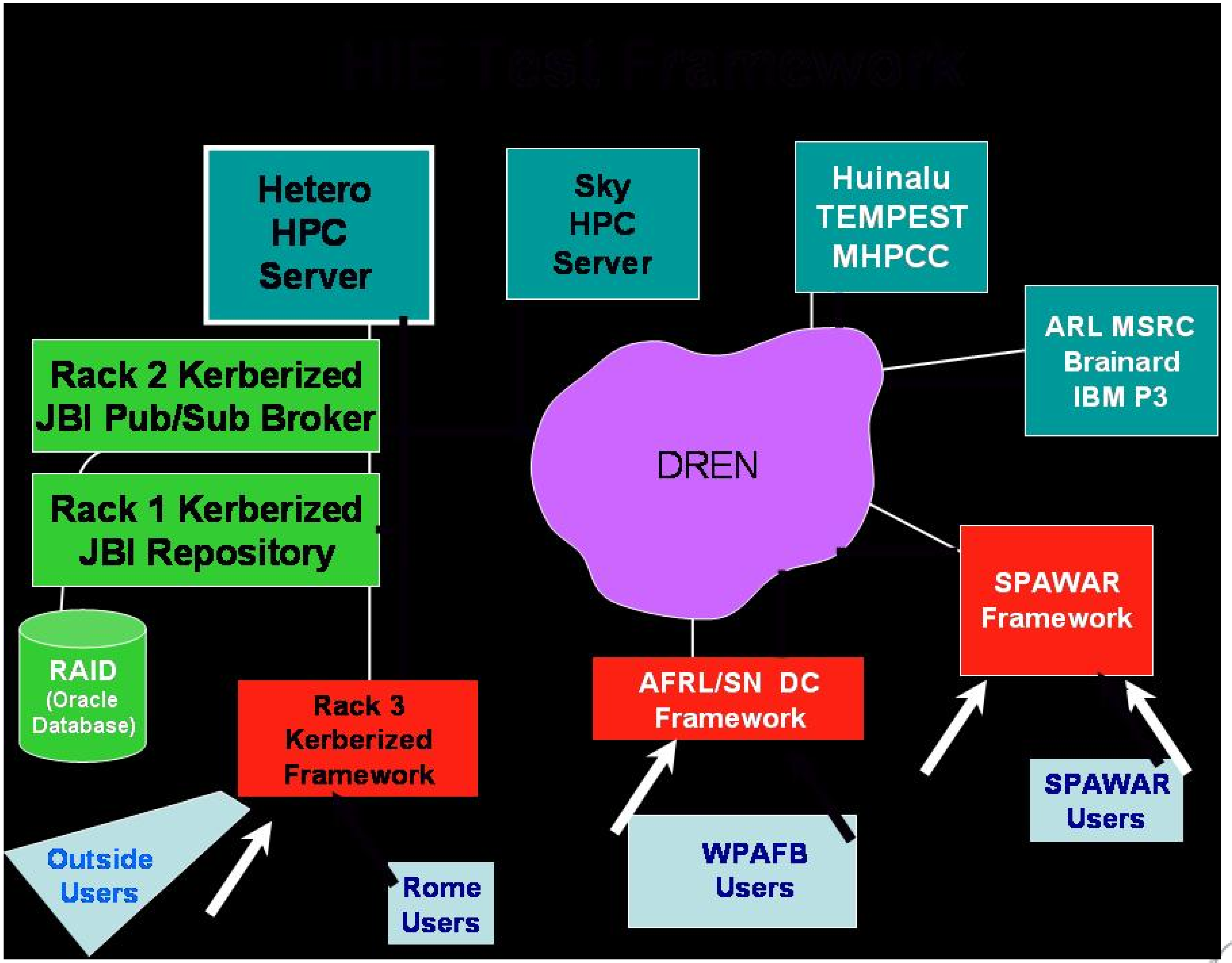
Distributed Applications

Hyperspectral Imaging Environment (HIE)



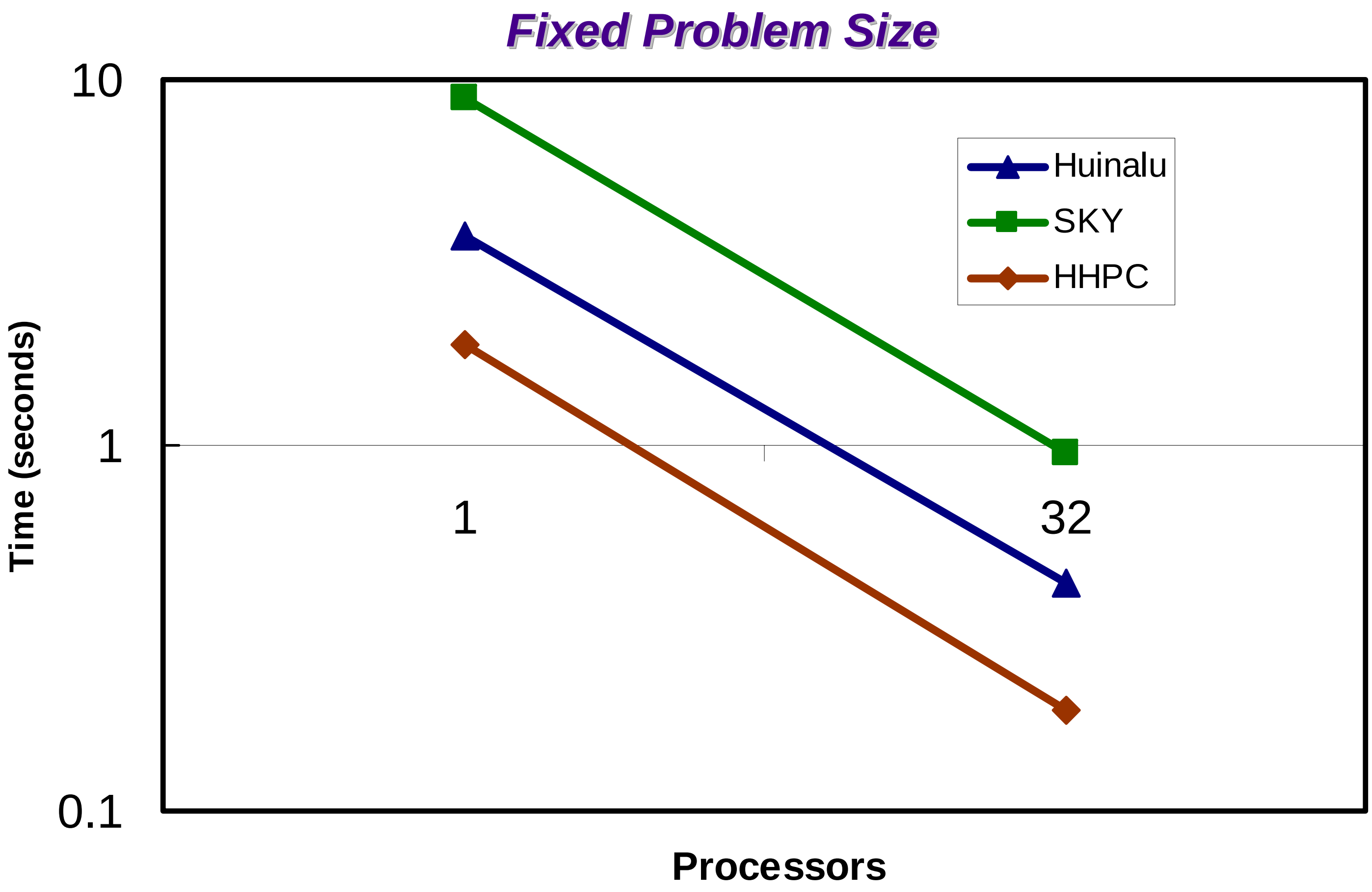
Electronic Battlefield Environment (EBE)

HIE Test Framework



MODTRAN

Processing Time v. Processors



Huinalu –
933 MHz P3

SKY –
333 MHz IBM

HHPC –
2.2GHz Intel Xeon

End Notes

- Challenges for Clusters:
  - Improve the robustness in a multi-use environment
  - Resolve porting issues
  - Compilers
  - Improve and mature the software environment
  - Improve system management tools

End Notes *(Continued)*

- Observations:
  - Current Cluster machines seem suited to jobs requiring less than 65 processors
  - If a job size approaches a significant fraction of the total system, instability increases
  - Clusters are "ready for prime time" for many applications but probably not for the more demanding scientific appliations